昆虫学报 *Acta Entomologica Sinica*, October 2013, 56(10): 1217－1228 ISSN 0454-6296

系统发育分析中的最大简约法及其优化

郑 巍[1,2,3]，罗阿蓉[2]，史卫峰[4]，郑为民[1,5]，朱朝东[2,*]
(1. 中国科学院深圳先进技术研究院，广东深圳 518055；2. 中国科学院动物研究所，动物进化与系统学(院)重点实验室，北京 100101；3. 中国科学院大学，北京 100049；4. 泰山医学院基础医学院，山东泰安 271016；5. 中国科学院信息工程研究所，北京 100093)

摘要：随着生物技术的不断发展和系统发育学的深入研究，在重构系统发育树时，研究人员往往要面对更多的挑战和困难，比如：(1)需要分析的样本数(物种数或个体数)不断增加；(2)需要分析的数据量迅速扩大。尤其在基因组测序技术的推动下，基于分子信息的系统发育重建需要极大的计算量，因此数学方法、计算机技术以及其他辅助工具对于系统发育重建的效率和精确度起着至关重要的作用。最大简约法(maximum parsimony)是一种重要的系统发育重建方法，提高其计算效率对系统发育学研究具有重要意义，针对该算法的优化改进需要生物学家和计算机专家的共同努力。本文通过详细地阐述最大简约法的计算流程，分析其参数选择对计算效率的影响，帮助更多的计算机使用者，在并不了解系统发育学基础的情况下，更方便地针对实际的系统发育算法问题给出更好、更快、更精准的解决方案；同时为系统发育研究工作者，较为清晰地解释最大简约法的构树思想和计算逻辑，推动针对最大简约法的不断改进与优化。

关键词：系统发育；系统发育重建；算法；最大简约法；计算流程；计算效率；优化
中图分类号：Q961 **文献标识码**：A **文章编号**：0454-6296(2013)10-1217-12

Phylogenetic algorithms: maximum parsimony and its optimization

ZHENG Wei[1,2,3], LUO A-Rong[2], SHI Wei-Feng[4], ZHENG Wei-Min[1,5], ZHU Chao-Dong[2,*] (1. Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China; 2. Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; 3. University of Chinese Academy of Sciences, Beijing 100049, China; 4. School of Basic Medical Sciences, Taishan Medical College, Tai'an, Shandong 271016, China; 5. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

Abstract: With the continuous development of biotechnoglogy and progresses in phylogenetics, researchers now are facing more and more challenges and difficulties in reconstructing phylogenetic trees: 1) species number (or individual number) of the specific taxon of research is always increasing; 2) the number of taxonomical characters (for example molecular information) of each species (or individual) is also enlarging. Especially with the efforts of genome-sequencing technology, phylogenetic reconstruction based on molecular information requires massive computation. Mathematical methods, computer technologies and other auxiliary means play key roles in enhancing the efficiency and accuracy of phylogenetic reconstruction. Maximum parsimony (MP) is a very important method for phylogenetic reconstruction, and it needs efforts of both biologists and computer scientists to enhance its computational efficiency. In this article, we elaborated the calculation procedure of the MP method in details and analyzed the influences of parameter selection on computational efficiency, in order to help more computer researchers without detailed knowledge of phylogenetics to present better, quicker and more precise solutions to phylogenetic reconstruction in practice. In the meantime, we tried to explain the basic principles and computational logic of the MP method for phylogenetic researchers to push forward continuous improvement and optimization of using maximum parsimony in biology.

基金项目：中国科学院知识创新工程重要方向项目(KSCX2-EW-B-02/03)；国家基础科学人才培养基金项目“特殊学科点”(J0930004, J1210002)；国家重大基础研究规划(“973”计划)项目(2006CB102003)；国家“十一五”科技支撑计划项目(2006BAD08A03)
作者简介：郑巍，男，1987 年 6 月生，黑龙江哈尔滨人，硕士研究生，研究方向为系统发育算法的优化，E-mail：zhengwei@ioz.ac.cn；wei.zheng@siat.ac.cn；zw870612@126.com
* 通讯作者 Corresponding author, E-mail：zhucd@ioz.ac.cn
收稿日期 Received：2013-03-15；接受日期 Accepted：2013-08-07

Key words: Phylogenetics; phylogenetic reconstruction; algorithm; maximum parsimony; calculation procedure; computational efficiency; optimization

针对系统发育的研究由来已久，早在亚里士多德时期，人们便开始对物种的性状进行描述。随着研究的深入，生物学家综合前人的理论、方法与成果，主要基于物种的基本形态性状，同时综合考虑其他行为、生理、生态和遗传等生物学差异，进行全面的系统发育研究，追溯物种的起源历史(黄大卫，1996)。然而在分子技术不断发展的帮助下，相比单纯地基于物种形态信息、物种性状信息进行系统发育研究，基于物种分子信息进行的系统发育研究如今更受推崇。一方面，分子信息能够有效替代物种的形态或性状信息，另一方面，分子信息可以作为物种进化的更为本质的内容来全面地描述物种、追溯历史。因此，通过分子信息进行系统发育研究，已经获得了绝大多数研究人员的认可，并在短短的几十年间，得到了迅速的发展和壮大(Suárez-Díaz and Anaya-Muňoz，2008)。

在基因组技术的推动下，基于基因组数据的系统发育分析，将帮助研究人员获得更为精准的分析结果。尽管研究表明单纯地通过增加基因数据并不能增强结果的一致性，只有配合复杂的分析研究，才能发挥全基因组的功能作用(Philippe *et al.*，2011)，但还是有越来越多的研究者开始使用全基因组数据进行系统发育重建，以期待获得更为全面的分析结果。在处理庞大的基因组数据时，研究人员不仅需要不断提高计算服务器的硬件配置，更需要迅速解决如何大幅度提高现有计算软件的分析效率这一难题(Delsuc *et al.*，2005)。因此，对大数据的系统发育重建软件的优化与改进，就显得尤为重要。

1 分子系统发育分析概述

1.1 分子系统发育重建的基本步骤和内容

一般来说，分子系统发育分析的主要步骤为(图 1)：生物学家通过采集标本，提取分子信息(如：DNA 序列等)，再将不同物种的分子序列信息进行多序列比对，获得可以统一比较的分子信息，通过计算机计算，输出最终的系统发育树。

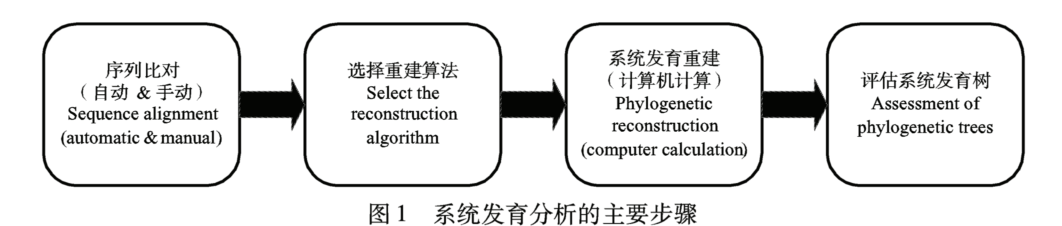


图 1 系统发育分析的主要步骤

Fig. 1 Main steps of phylogenetic analysis

作为系统发育研究的基础和内容，以及后续计算的输入来源，“分子水平上的生物信息获取”显得尤为重要。一般来说，分子信息主要由 DNA、RNA 和蛋白质三大类生物分子信息组成，并从 3 个方面进行信息提取：分子序列信息、分子空间结构信息、分子功能性状信息。分子空间结构相对复杂，目前还没有较为统一的、有效的研究标准(Lin and Gerstein，2000)。分子功能性状往往受到外界环境的影响，很难真实地反映物种的进化特性。结合分子二级结构与分子序列双重信息进行比对与重构系统发育树的研究还在实验阶段，难度较大(Letsch *et al.*，2010)。因此，目前较为常用的方法是从物种(或个体)获得部分(或全部)分子序列信息，通过特定的研究方法和手段，结合现代数理统计与计算机科学技术，重构物种(或个体)之间的生物系统发育关系(Philippe *et al.*，2005；张树波和赖剑煌，2010)。

通常情况下，“多序列比对”与“系统发育重建”是分先后依次进行的。然而由于二者的计算任务都相当耗时，且计算内容有重叠部分，因此目前也有研究尝试将二者结合(Roshan *et al.*，2006)，共同计算，以期待获得更高的计算效率，降低计算成本。在这一方面，最大简约法的特性尤为突出，研究人员证实，最大简约法能够帮助“多序列比对”和“系统发育重建”共同进行(Liu *et al.* ，2009)。

明确了所研究问题的输入信息后，就需要进一步选择系统发育重建算法。一般来说，基于分子序列信息的系统发育重建算法主要可分为两大类：基

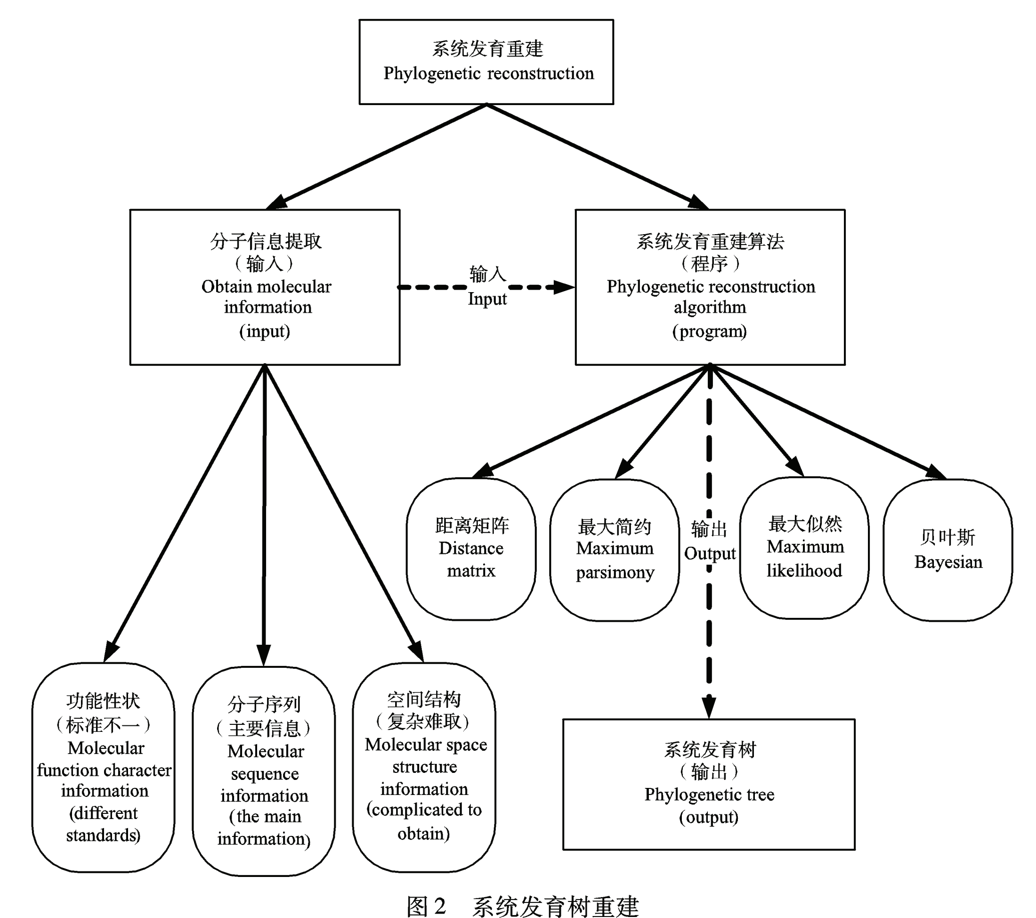


图 2　系统发育树重建

Fig. 2　Reconstruction of phylogenetic tree

于距离矩阵的系统发育重建算法（Fitch and Margoliash，1967；Kidd and Sgaramella-Zonta，1971）和基于最优原则的系统发育重建算法（李建伏和郭茂祖，2006）。基于最优原则的方法主要又可分为最大简约法（Fitch，1971）、最大似然法（Felsenstein，1981）以及贝叶斯法（Huelsenbeck and Ronquist，2001）（Yang and Rannala，1997）3 种算法。

1.2　分子系统发育重建的难点与优化

在一个含有 n 个物种（或个体）的分类单元中，存在(2n－3)!! 种可能的有根树拓扑结构，(2n－5)!! 种可能的无根树拓扑结构（Roch，2006）。因此计算机在“树空间”中搜索最优系统发育树（如：最大简约树）是一个 NP（non-deterministic polynomial，非确定多项式）难问题（Foulds and Graham，1982），人们只能通过一些近似假设和算法优化设计，获得最优近似解，作为最终输出的系统发育树。因此，算法上的改进是系统发育重建效率提升的关键所在。

以最大简约法进行系统发育重建为例，计算机算法上的优化问题，概括来说，主要体现在“树空间”搜索和“最大简约值”计算两个步骤上（图 3）。一方面，在系统发育重建问题中，采用穷举式搜索属于 NP 难问题（Day *et al.*，1986），需要使用“启发式搜索”进行近似求解；另一方面，基因组的庞大数据，要求“最大简约值”的计算效率不断提高，以适应不断扩充的输入数据。

2　最大简约法概述

最大简约法（maximum parsimony）简称为 MP 法，最早源于形态性状研究，现在已经推广到分子序列的进化分析中。最大简约法认为对于一个分类群来说，所有可能的系统发育树中，性状或基因变化总和最小的那一棵系统发育树是真正接近自然变化的系统发育拓扑（Henning，1966）。换句话说，当给定一个物种类群之后，每个个体的性状或分子序列便已知，研究人员事先根据实际生物学含义规定出各个性状或基因相互转化和突变的代价大小，从而针对每一棵可能的系统发育树进行分析，计算其总共的转化和突变的代价大小，最终选择代价最

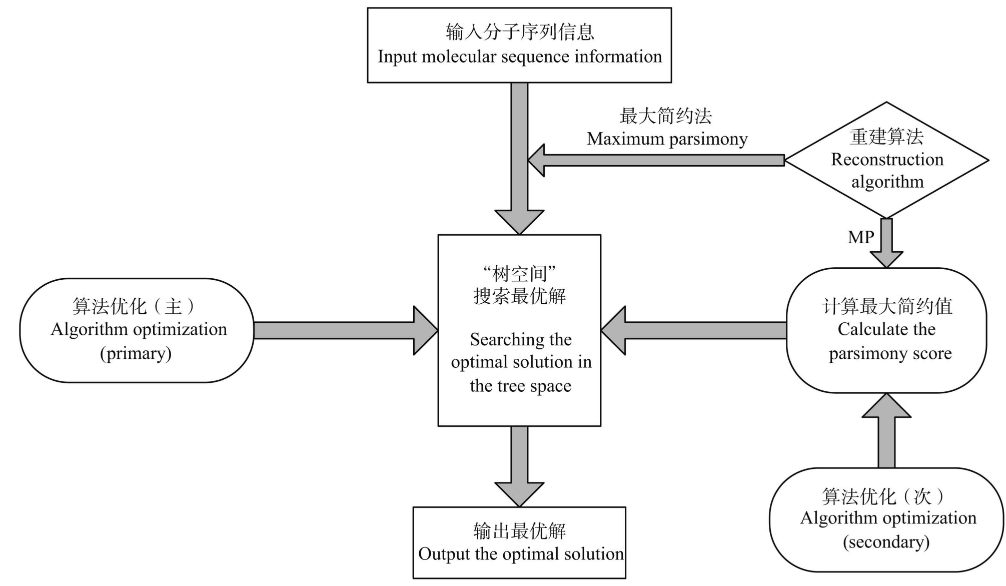


图 3 基于最大简约法重构系统发育树的算法优化

Fig. 3 Algorithm optimization of phylogenetic reconstruction by Maximum Parsimony

小的一棵树为这个物种类群的 MP 树(最大简约树)(Felsenstein, 1983)。

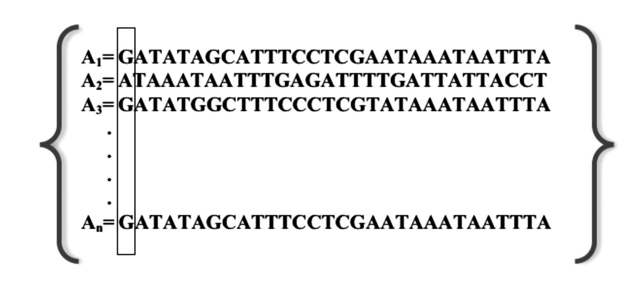
在系统发育重建算法中，最大简约原则有着重要的地位和影响力，它最早由 Camin 等人提出(Camin and Sokal, 1965)，随后 Hein 等人对其如何重建祖先状态(Hein, 1990)以及构建系统发育树，进行了深入研究与推广(Hein, 1993)。一般来说，系统发育重构时所需要的历史进化信息越少，则其所得结果就越发真实可信(Sober, 1988)，而较早出现的最大简约法又恰恰是一种不需要进化模型的无噪声统计方法(Sourdis and Nei, 1988)；且在估算单倍型基因迁移的研究中，无论是从结果的精确度还是在算法的鲁棒性上，最大简约法相对于最大似然法、距离矩阵法、贝叶斯法都略胜一筹(Salzburger *et al.*, 2011)；同时，在其他系统发育重建方法(Brooks *et al.*, 2007)和系统发育网络研究(Jin *et al.*, 2006)中都或多或少使用了简约原则，因此最大简约法有着深厚的、广泛的系统发育学者的认同。

3 最大简约法的计算步骤

3.1 获取输入信息

最大简约法最早是以物种的形态学性状作为分析内容，进行比较计算(Hein, 1990)，重建物种间的系统发育树。随着分子技术的不断发展，如今最大简约法主要针对 DNA、RNA 和蛋白质分子进行分析研究。以 n 条 DNA，每条 DNA 中包含 30 个脱氧核苷酸位点的输入信息为例，n 个物种的 DNA 经过序列比对，形成了含有 30 * n 个位点信息的 DNA "矩阵"，作为最大简约法计算的输入信息。

在实际计算中，最大简约法每次只针对 n 个物种的一个位点信息进行计算。如图 4 所示，抽取第 1 个位点信息列。计算结束后，再抽取第 2 个位点信息列进行计算，以此类推，直到最后一个位点完成计算。



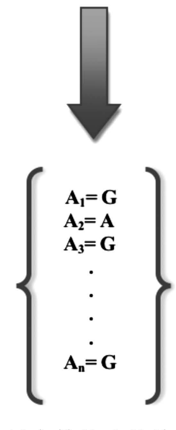


图 4 最大简约法的输入信息

Fig. 4 The input files for maximum parsimony

3.2 “树空间”搜索最大简约树

获得输入信息后，针对 n 个物种的每一列位点信息，在“树空间”中搜索最大简约值最小的系统发育树。由于在“树空间”搜索最优解是一个 NP 难问题(Day，1987)，对于物种数 n 大于 10 的分类单元，一般只能采用启发式智能搜索，获得最优近似解作为结果。

其具体做法是：首先，按照“逐步添加算法(stepwise addition algorithm)”(Cavalli-Sforza and Edwards，1967)或“星状分解算法(star decomposition algorithm)”生成一棵系统发育树(称之为“初始树”)，随后采用“邻居互换法(nearest-neighbour interchange，NNI)”或“子树修剪与重接法(subtree pruning and regrafting，SPR)”或“树对切与重接法(tree bisection and reconnection，TBR)”对其进行反复修正，直到获得该位点信息下的 n 个物种的最大简约树(图 5)。

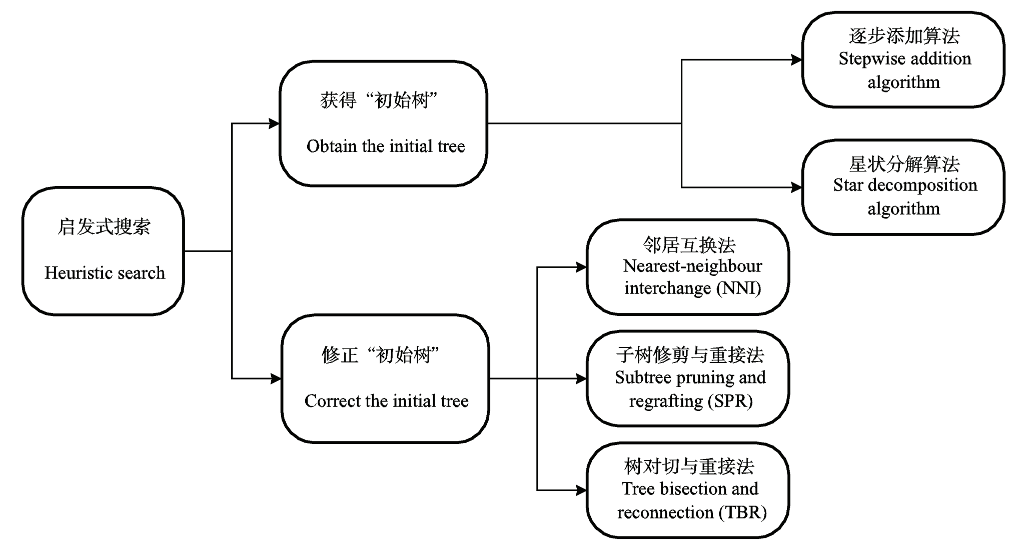


图 5　启发式搜索算法

Fig. 5　Heuristic search algorithm

3.3 最大简约值(函数值)的评价标准与计算方法

除了在庞大的“树空间”搜索最优解这一难题外，“最大简约值”的计算也是影响算法效率的关键所在。通过计算“最大简约值”得到两棵树之间优劣比较的评判标准。一般来说，对于一棵特定拓扑结构的系统发育树，最大简约值的计算过程如图 6 所示。通过对特定拓扑结构的系统发育树的祖先情况进行穷举式的反复推断，计算出各种推断可能的进化“代价”，选择进化“代价”最小的推断情况，作为该棵树的祖先信息，同时以其进化“代价”作为该棵树的“最大简约值”(函数值)。对“树空间”的每一棵树的“函数值”进行比较，其中“函数值”最小的一棵树作为该位点信息列的最大简约树(MP 树)(图 7)。

然而，实际操作中，程序并不是对每一个祖先情况进行计算，也并非要计算出所有的系统发育树的“最大简约值”(函数值)。正如前文所说，由于“树空间”的解的个数，随着物种数 n 的增加呈指数式扩增，同时祖先推断的选取也是一个 NP 难问题(Bader *et al.*，2006)，因此，人们一方面在祖先推断中，采用动态规划算法(dynamic-programming algorithm)(Sankoff，1975)进行树长计算(“最大简约值”的计算)；另一方面，采用启发式搜索，对“树空间”进行搜索扫描，以最快的方式，精准地获得近似的最大简约树(Yang，2006)。

3.4 全序列的最大简约树

通过上述步骤，可以获得每一列位点信息的最大简约树，针对 30 个位点长度的 DNA 序列“矩阵”，就得到了 30 棵基于单位点信息的最大简约树。采用频率法、树长比较法、改进频率法等方法，来获得最终的最大简约树(全局 MP 树)。

3.5 重采样过程

容易看出，上述过程获得的最大简约树，并不具备统计学意义；同时由于 DNA 序列过于复杂，本身采用近似算法得到的结果不足以让研究人员所信服，因此，人们通过重采样过程，对上述过程获得

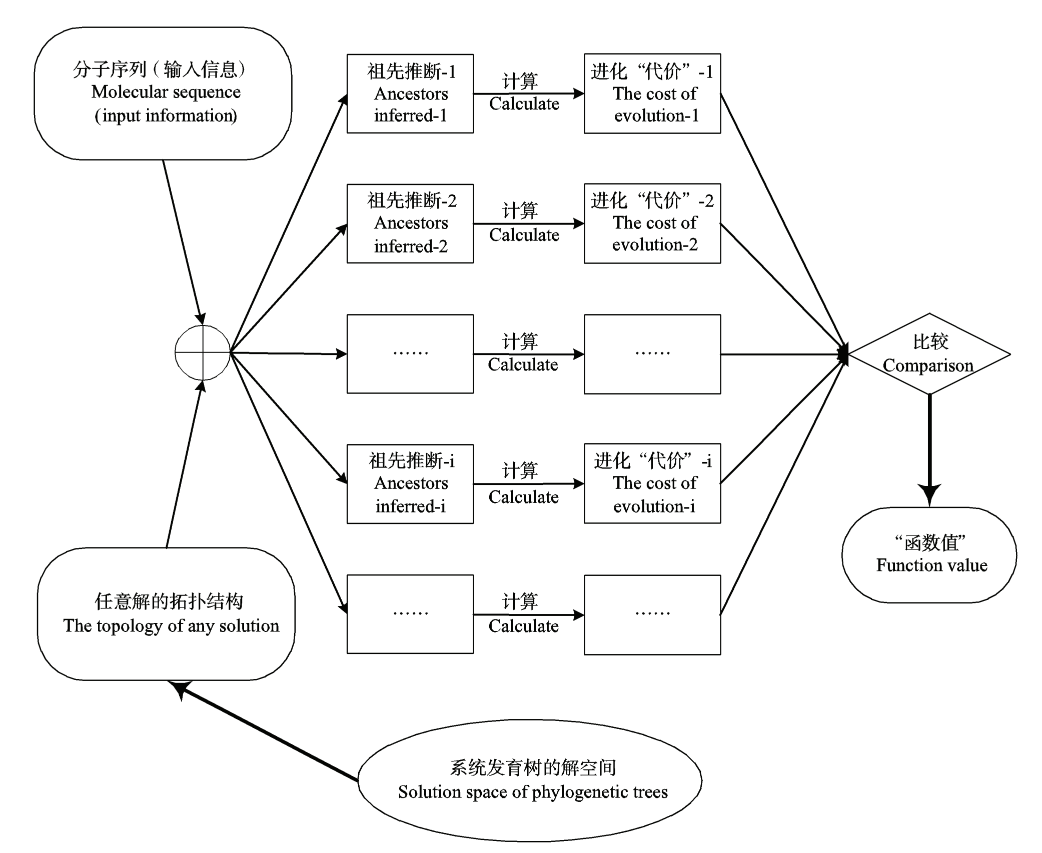


图 6　系统发育树的“函数值”计算过程

Fig. 6　Function value calculation process of phylogenetic tree

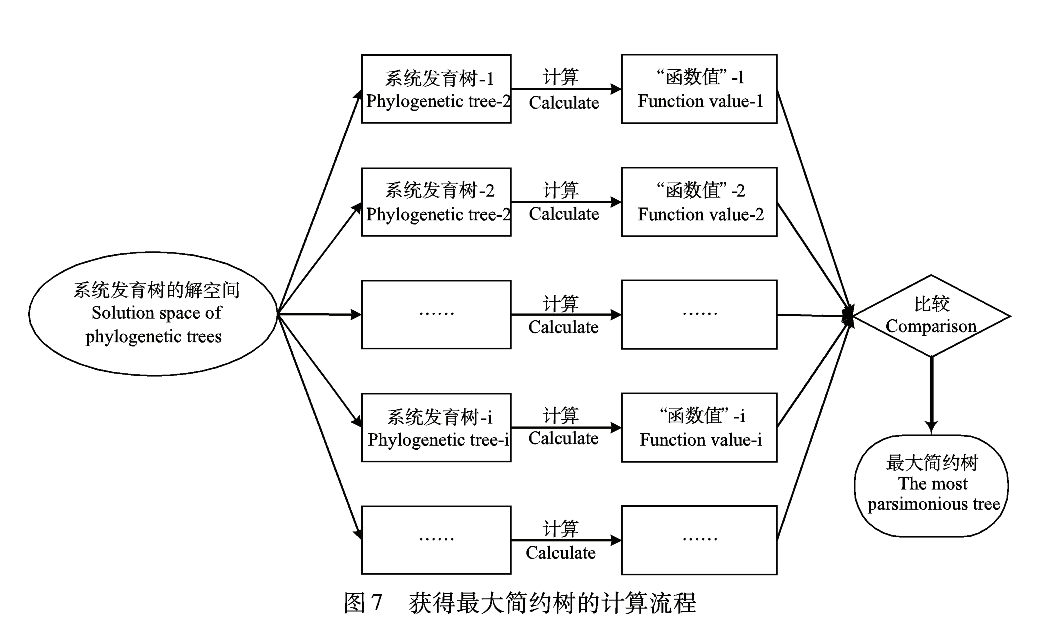


图 7　获得最大简约树的计算流程

Fig. 7　Calculation procedure of obtaining the most parsimonious tree

的最大简约树进行评估，来证明最大简约树的正确性以及精确度（Felsenstein，1985）。具体做法是：对原有 DNA“矩阵”进行随机地、可重复地抽样（Stamatakis *et al.*，2008），形成新的“大小”与原“矩阵”相同的抽样输入信息（图 8），重复计算，获得新的最大简约树，如此往复成千上万次，得到大量的抽样最大简约树，进而与最开始的最大简约树比较，获得支持率（Bhattacharya，1996），使其具有统计学意义上的支持（Alfaro *et al.*，2003），以证明其正确性与精确度。

DNA序列矩阵（输入）
DNA sequence matrix (input)

重采样分析
Bootstrap

重采样10 000次
Resampling 10 000

位点1信息
Information of site 1

位点2信息
Information of site 2

位点3信息
Information of site 3

……

位点M信息
Information of site M

获得“初始树”
（逐步添加法、星状分解法）
Obtain the initial tree
(stepwise addition algorithm, star decomposition algorithm)

计算树的“最大简约值”
Calculate the trees' parsimony score

修正“起始树”
（NNI, SPR, TBR）
Correct the initial tree
(NNI, SPR, TBR)

重复左侧计算流程
Repeat the calculation process of the left side

位点1“MP树”
MP tree of site 1

位点2“MP树”
MP tree of site 2

位点3“MP树”
MP tree of site 3

……

位点M“MP树”
MP tree of site M

获得最优MP树
Obtain optimal MP tree

统计分析
Statistical analysis

10 000 棵重采样 MP 树
10 000 resampling MP trees

最终输出MP树（带支持率）
The final output of the MP tree
(with the confidence rate)

图 9　基于最大简约原则的分子系统发育重建计算流程图

Fig. 9　The calculation flowchart of molecular phylogenetic reconstruction based on the principle of maximum parsimony

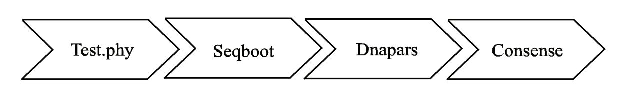


图 10　Phylip 软件包基于最大简约法的使用流程

Fig. 10　The using process of Phylip packages based on maximum parsimony

耗时可以忽略不计。因此，本文在后续分析中，仅考虑 Dnapars 时间部分，以便于研究比较。

4.4　“Number of trees to save”的直接影响

在 Dnapars 分析中，选项“number of trees to save”表示在计算结果中保留等价系统发育树的个数，save 值越大，表明程序搜索解集的范围越大，所得结果越可信，所需计算时间越长。在通过对不同 save 值的测试中(表 2)，发现计算时间的增长比例，同 save 值的增长比例在同一个数量级上，且基本相等。因此，为了获得更为精确的分析结果，提

表 1　Seqboot/Dnapars/Consense 分析过程的耗时(s)

Table 1　Time-consuming (s) of Seqboot/Dnapars/Consense

数据大小 Species-Sites	Seqboot (重采样 100 次) (replicates = 100)	Dnapars (无重采样分析) (no replicates)	Consense (无重采样分析) (no replicates)	Dnapars 与 Seqboot 分析过程的时间比 Dnapars: Seqboot	Dnapars 与 Consense 分析过程的时间比 Dnapars: Consense
56 – 24	0.00	0.02	0.00	–	–
56 – 249	0.04	0.39	0.00	9.75	–
56 – 540	0.08	2.35	0.00	29.38	–
56 – 1 197	0.20	3.26	0.00	16.30	–
357 – 24	0.02	14.01	0.02	700.50	700.50
357 – 249	0.28	35.94	0.02	128.36	1 797.00
357 – 540	0.61	106.71	0.01	174.93	10 671.00
357 – 1 197	1.70	389.42	0.02	229.07	19 471.00
962 – 24	0.08	899.09	0.23	11 238.63	3 909.09
962 – 249	0.78	1 087.31	0.28	1 393.99	3 883.25
962 – 540	1.61	2 425.46	0.37	1 506.50	6 555.30
962 – 1 197	4.81	6 531.91	0.35	1 357.99	18 662.60

表 2　Dnapars 计算耗时(s)随 save 值的增长——成倍增长

Table 2　Time-consuming (s) of Dnapars increases with the growth of the save value exponentially

数据大小 Species-Sites	save 值为 10 Save = 10	save 值为 100 Save = 100	save 值为 1 000 save = 1 000	Save 为 100 与 10 时，Dnapars 分析的时间比 Save 100:10	Save 为 1 000 与 100 时，Dnapars 分析的时间比 Save 1 000:100	Save 为 1 000 与 10 时，Dnapars 分析的时间比 Save 1 000:10
56 – 24	0.02	0.02	0.02	1.00	1.00	1.00
56 – 249	0.39	3.44	32.38	8.82	9.41	83.03
56 – 540	2.35	18.20	147.41	7.74	8.10	62.73
56 – 1197	3.26	25.37	234.43	7.78	9.24	71.91
357 – 24	14.01	74.13	697.87	5.29	9.41	49.81
357 – 249	35.94	365.19	2 741.86	10.16	7.51	76.29
357 – 540	106.71	1 008.64	18 553.95	9.45	18.40	173.87
357 – 1 197	389.42	2 902.94	33 843.48	7.45	11.66	86.91
962 – 24	899.09	4 867.84	64 365.70	5.41	13.22	71.59
962 – 249	1 087.31	6 820.49	70 058.93	6.27	10.27	64.43
962 – 540	2 425.46	27 653.34	30 0038.74	11.40	10.85	123.70
962 – 1 197	6 531.91	31 154.97	38 9 748.67	4.77	12.51	59.67

高 m 倍的 save 值，将带来近 m 倍的时间成本的增长。

4.5　重采样分析的耗时

重采样分析过程相当于将一次 Dnapars 分析重复进行成百上千次，尽管输入数据经过重新抽样已经不同于原始输入数据，但是数据大小并没有改变，因此重采样过程的计算时间将会在原数据基础上成倍线性增长，耗时巨大。本文对不同大小的数据分别在 replicates = 1，10，50 和 100 的情况下计算(replicates = 1 表示只进行原始数据的 Dnapars 分析，不进行重采样数据的 Dnapars 分析)，从耗时数据(表 3)来看，容易发现：(1) replicates = 1 时，存在较大的偶然性偏差(由程序本身算法结构造成)，尽管耗时随着数据增大而增大，但计算时间随着

replicates 增长时，并没有呈线性正比例关系增长；(2)在 replicates = 10，50 和 100 的实验结果比较中，能够发现耗时的增长比例，基本等于 replicates 的增长比例。这主要是因为随着 replicates 的增长，多组数据的重复实验，消除了 replicates = 1 情况下单组数据的计算偶然性因素，真实地、平均地反映了程序计算量的大小。

表 3 Replicates = 1，10，50 和 100 时的 Dnapars 耗时(s)

Table 3 Time-consuming (s) of Dnapars while replicates equals to 1，10，50 and 100

	数据大小 Species-Sites							
	56 – 24	56 – 249	56 – 540	56 – 1 197	357 – 24	357 – 249	357 – 540	357 – 1 197
重采样分析 1 次 Replicates = 1	0.02	0.39	2.35	3.26	14.01	35.94	106.71	389.42
重采样分析 10 次 Replicates = 10	16.66	56.95	194.42	292.60	7 062.62	11 604.70	26 121.23	40 873.46
重采样分析 50 次 Replicates = 50	82.62	296.60	932.09	1 360.73	34 900.50	57 026.15	110 559.02	242 711.08
重采样分析 100 次 Replicates = 100	170.17	599.02	1 871.30	2 716.93	67 091.81	113 221.90	205 704.82	425 106.68
重采样分析 10 次与 1 次的耗时比 Replicates 10:1	833.00	146.03	82.73	89.75	504.11	322.89	244.79	104.96
重采样分析 50 次与 1 次的耗时比 Replicate 50:1	4 131.00	760.51	396.63	417.40	2 491.11	1 586.70	1 036.07	623.26
重采样分析 100 次与 1 次的耗时比 Replicate 100:1	8 508.50	1 535.95	796.30	833.41	4 788.85	3 150.30	1 927.70	1 091.64
重采样分析 50 次与 10 次的耗时比 Replicate 50:10	4.96	5.21	4.79	4.65	4.94	4.91	4.23	5.94
重采样分析 100 次与 10 次的耗时比 Replicate 100:10	10.21	10.52	9.63	9.29	9.50	9.76	7.88	10.40
重采样分析 100 次与 50 次的耗时比 Replicate 100:50	2.06	2.02	2.01	2.00	1.92	1.99	1.86	1.75

4.6 并行化重采样过程

3.6 节中总结的算法结构显示，重采样分析中各个 DNA“矩阵”作为输入信息，在计算过程里是相互独立的，互不影响；4.5 节的计算效率分析同样表明，replicates 越小，Dnapars 的计算时间成本越低。据此，针对重采样分析过程，作者设计了并行化计算结构，将大量重复的 Dnapars 计算分布在不同的计算核心上并行执行。本文对不同大小的真实数据，分别在单核上串行计算，双核和四核上并行计算，其结果(表4)表明：(1)在并行环境下，基于最大简约法的系统发育重建效率得到了大大提升；(2)双核并行将计算时间缩短到单核串行时间的近二分之一，四核并行缩短到近四分之一；(3)在硬件服务器配置完全满足的情况下，并行计算重采样过程，能够有效提高分析效率，帮助生物学家更好地进行分析研究。

表 4　并行执行重采样分析的时间消耗

Table 4　Time-consuming of bootstrap process using parallel computing

数据大小 Species-Sites	耗时 Time-consuming (s)			耗时比 Ratio of time-consuming		
	单核串行 Serial 1-core	双核并行 Parallel 2-core	四核并行 Parallel 4-core	双核并行: 单核串行 2-core: 1-core	四核并行: 双核并行 4-core: 2-core	四核并行: 单核串行 4-core: 1-core
56 – 24	170.17	86.00	43.35	0.51	0.50	0.25
56 – 249	599.02	297.23	148.40	0.50	0.50	0.25
56 – 540	1 871.30	937.62	474.75	0.50	0.51	0.25
56 – 1 197	2 716.93	1 372.35	674.44	0.51	0.49	0.25
357 – 24	67 091.81	34 936.54	16 723.05	0.52	0.48	0.25
357 – 249	113 221.90	57 129.17	27 982.24	0.50	0.49	0.25
357 – 540	205 704.82	111 138.64	58 146.46	0.54	0.52	0.28
357 – 1 197	425 106.68	250 152.30	128 748.97	0.59	0.51	0.30

5　讨论与展望

随着基因组时代的到来，针对最大简约法的现有计算软件已经无法满足生物学家的研究需求，一方面，分析的生物类群中所含的个体数越来越多；另一方面，每个个体的分子信息含量不断增长，这些无疑将带来巨大的计算量。因此，针对最大简约法在实际使用中的改进与优化，将变得任重而道远。单纯地依靠计算服务器并行化 bootstrap 过程，对于最大简约法提高效率，是治标不治本的。在计算效率提升方面，作者认为至少还有以下几点可以考虑：(1)最大简约值的计算过程的并行化设计；(2)“树空间”的搜索方法(启发式搜索)的智能化改进；(3)基于类似蚁群算法、遗传算法改进启发式搜索过程后的并行化改进。

当然，这些仅仅是针对最大简约法改进的一个开始，作者在此希望通过系统地、概括地剖析最大简约算法，能够为有志于此的计算机工作者提供些许帮助，以不断推动最大简约法的计算优化与改进，进一步满足生物学工作者的计算需求。

参考文献(References)

Alfaro ME, Zoller S, Lutzoni F, 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.*, 20(2): 255 – 266.

Bader DA, Chandu VP, Yan M, 2006. ExactMP: an efficient parallel exact solver for phylogenetic tree reconstruction using maximum parsimony. In: Proceeding of the 35th International Conference on Parallel Processing (ICPP2006) in Columbus, OH. 65 – 73.

Bhattacharya D, 1996. Analysis of the distribution of bootstrap tree lengths using the maximum parsimony method. *Molecular Phylogenetics and Evolution*, 6(3): 339 – 350.

Brooks DR, Bilewitch J, Condy C, 2007. Quantitative phylogenetic analysis in the 21st century. *Revista Mexicana de Biodiversidad*, 78: 225 – 252.

Bryant D, 2003. A classification of consensus methods for phylogenetics. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 61: 163 – 184.

Camin JH, Sokal RR, 1965. A method for deducing branching sequences in phylogeny. *Evolution*, 19(3): 311 – 326.

Cavalli-Sforza LL, Edwards AWF, 1967. Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics*, 19: 233 – 257.

Day WHE, 1987. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49(4): 461 – 467.

Day WHE, Johnson DS, Sankoff D, 1986. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81(1): 33 – 42.

Delsuc F, Brinkmann H, Philippe H, 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6: 361 – 375.

Felsenstein J, 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6): 368 – 376.

Felsenstein J, 1983. Parsimony in systematics: biological and statistical issues. *Ann. Rev. Ecol. Syst.*, 14: 313 – 333.

Felsenstein J, 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4): 783 – 791.

Fitch WM, 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Biol.*, 20 (4): 406 – 416.

Fitch WM, Margoliash E, 1967. Construction of phylogenetic trees. *Science*, 155(3760): 279 – 284.

Foulds LR, Graham RL, 1982. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3(1): 41 – 49.

Hein J, 1990. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Biosciences*, 98(2): 185 – 200.

Hein J, 1993. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36(4): 396 – 405.

Henning W, 1966. Phylogenetic Systematics. University of Illinois Press, Urbana.

Huang DW, 1996. An Introduction to Cladistics. China Agriculture Press, Beijing. 1 – 10. [黄大卫, 1996. 支序系统学概论. 北京: 农业出版社. 1 – 10]

Huelsenbeck J, Ronquist F, 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8): 754 – 755.

Jin G, Nakhleh L, Snir S, Tuller T, 2006. Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics*, 23(2): e123 – e128.

Kidd KK, Sgaramella-Zonta LA, 1971. Phylogenetic analysis: concepts and methods. *Am. J. Hum. Genet.*, 23(3): 235 – 252.

Letsch HO, Kück P, Stocsits RR, Misof B, 2010. The impact of rRNA secondary structure consideration in alignment and tree reconstruction simulated data and a case study on the phylogeny of hexapods. *Mol. Biol. Evol.*, 27(11): 2507 – 2521.

Li JF, Guo MZ, 2006. A review of phylogenetic tree reconstruction technology. *Acta Electronica Sinica*, 34(11): 2047 – 2052. [李建伏, 郭茂祖, 2006. 系统发生树构建技术综述. 电子学报, 34(11): 2047 – 2052]

Lin J, Gerstein M, 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.*, 10: 808 – 818.

Liu K, Nelesen S, Raghavan S, Linder CR, Warnow T, 2009. Barking up the wrong treelength: the impact of gap penalty on alignment and tree accuracy. *IEEE TCBB*, 6(1): 7 – 21.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D, 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.*, 9(3): e1000602.

Philippe H, Delsuc F, Brinkmann H, Lartillot N, 2005. Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.*, 36: 541 – 562.

Roch S, 2006. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE TCBB*, 3(1): 92.

Roshan U, Livesay DR, Chikkagoudar S, 2006. Improving progressive alignment for phylogeny reconstruction using parsimonious guide-trees. In: Proceeding of the IEEE 6th Symposium on Bioinformatics and Bioengineering (BIBE06), Washington D. C. 159 – 164.

Salzburger W, Ewing GB, Von Haeseler A, 2011. The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Molecular Ecology*, 20(9): 1952 – 1963.

Sankoff D, 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, 28(1): 35 – 42.

Sober, 1988. Reconstructing the Past: Parsimony, Evolution and Inference. MIT Press, Cambridge.

Sourdis J, Nei M, 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Mol. Biol. Evol.*, 5(3): 298 – 311.

Stamatakis A, Hoover P, Rougemont J, 2008. A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.*, 57(5): 758 – 771.

Suárez-Díaz E, Anaya-Muñoz VH, 2008. History, objectivity, and the construction of molecular phylogenies. *Stud. Hist. Phil. Biol. Biomed. Sci.*, 39(4): 451 – 468.

Yang Z, Rannala B, 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol.*, 14(7): 717 – 724.

Yang ZH, 2006. Computational Molecular Evolution. Oxford University Press, London. 82 – 90.

Zhang SB, Lai JH, 2010. Bioinformatics approach for molecular evolution research. *Computer Science*, 37(8): 47 – 51. [张树波, 赖剑煌, 2010. 分子系统发育分析的生物信息学方法. 计算机科学, 37(8): 47 – 51]

(责任编辑: 袁德成)